



AI Pentest

Managed AI penetration testing with real scope control, isolated execution, and customer-ready reports.

Product presentation

Security testing is hard to scale safely

Teams need broader coverage without losing authorization, evidence, or operational control.

Manual bottleneck

Expert time is spent on repetitive recon, endpoint review, and report assembly before judgment begins.

Tool sprawl

Scanners, browser traces, notes, PoCs, and status updates are often disconnected.

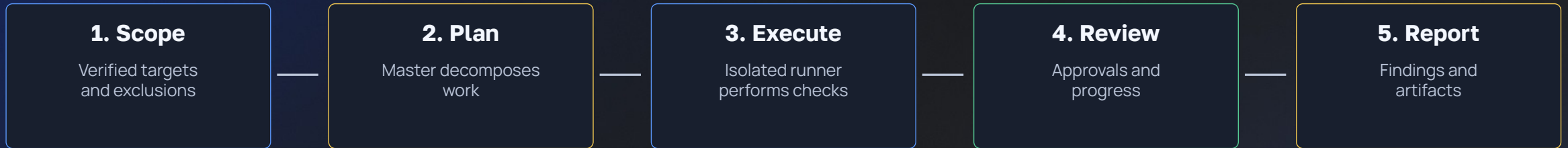
Unsafe autonomy

Unbounded agents can cross scope, hide decisions, or create traffic nobody approved.

The product goal: make AI useful for pentest delivery while every action remains accountable.

Triadix turns AI pentesting into a managed run

Scope, orchestration, review, evidence, and reporting stay inside one platform workflow.

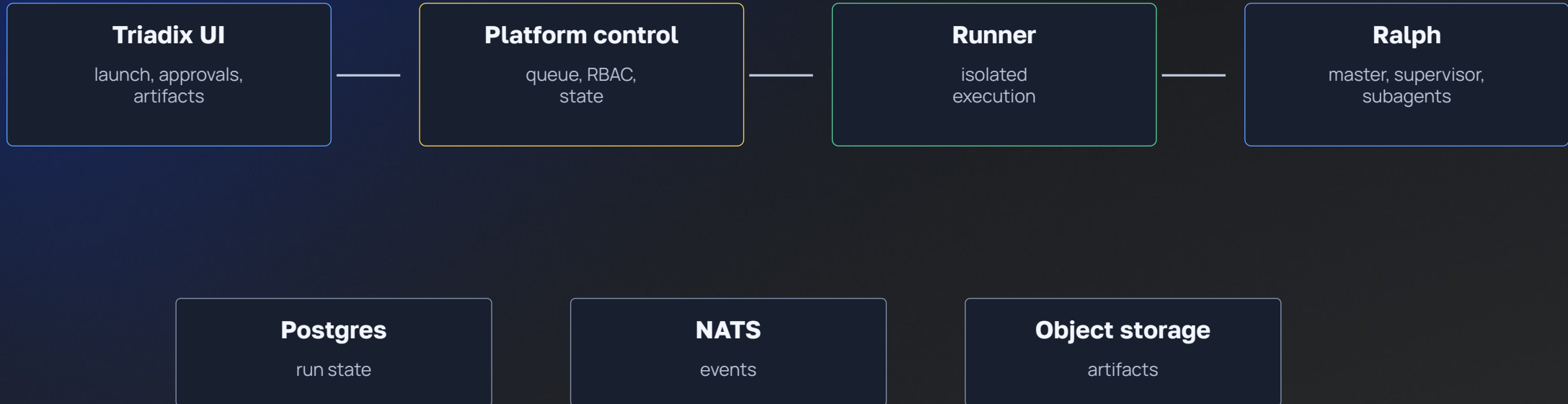


One sentence

Triadix is the control plane for AI pentest delivery; the runner is the isolated execution plane that performs work and reports back.

Control plane and execution plane are separated

The platform remains the source of truth while target-facing work runs in isolation.



No direct browser-to-runner path. Runners claim work outbound and sync structured snapshots back.

Ralph coordinates the multi-agent run

A master agent delegates narrow work to subagents and reviews their artifacts.



Key design choice: subagent work is visible through reviewable artifacts, not hidden agent memory.

Operator workflow

The user sees a controlled assessment lifecycle, not an opaque autonomous bot.



Human checkpoints

Pause for scope ambiguity, high-risk actions, credentialed testing, state-changing checks, or enough evidence for handoff.

Two orchestration modes

Mode controls how subagents are allocated across authorized scope.

Deep Focus

Deep testing for narrow scope with aggressive pursuit of important findings.

Best for: one app, one API, one known surface, or a specific hypothesis.

Broad Coverage

Wide coverage for large or unclear scope with soft branch balancing.

Best for: many hosts, broad web estates, external perimeter, or early discovery.

Both modes preserve the same approval, isolation, and reporting model.

Safety is a product feature, not a prompt

The runtime combines scope policy, approvals, isolation, and auditability.

- Verified scope and immutable launch context
- Human approval for sensitive or high-risk actions
- Runtime auto-approve can be toggled during a run
- Reports and artifacts go through platform permissions
- Raw prompts, secrets, and internals stay out of user output
- Workers run under a dedicated non-root UID
- Worker egress blocks platform internals and control networks
- Controller keeps only the control-plane access it needs
- Private targets require explicit per-run allow rules
- Resource limits cover CPU, memory, PIDs, logs, and scratch space

AI pentest workspace

Runs, findings, and report artifacts are visible from one operational screen.

The screenshot shows the TRIADIX AI pentest workspace. The left sidebar contains navigation links: Dashboard, AI pentest (selected), Projects, Calendar, Attack Surface, Diagnostics, and Organization. The main content area is titled 'AI pentest' and includes a 'New pentest' button. Below this is a table of 'Pentest runs' with columns for ID, Target, Mode, Status, Stage, Depth, Report, and Date. Two runs are listed, both with a status of 'Succeeded'.

ID	TARGET	MODE	STATUS	STAGE	DEPTH	REPORT	DATE
#892be1d0	*.ctfonly.fans	Deep	Succeeded	Report	Standard	Report	Jul 04, 2026, 04:33 PM
#5a82d374	tensor.ctfonly.fans	Deep	Succeeded	Report	Standard	Report	Jul 03, 2026, 11:11 PM

What it shows

- Run history
- Status and stage
- Depth and mode
- Report readiness

Run overview

Operators can see the current stage, progress, and ready artifacts at a glance.

The screenshot displays the 'Run overview' interface for an AI pentest project. The left sidebar shows the 'Triadix Production' environment and a navigation menu with options like Dashboard, AI pentest, Projects, Calendar, Attack Surface, Diagnostics, and Organization. The main content area is titled 'CUSTOMER WORKSPACE AI pentest' and shows the project name 'tensor.ctfonly.fans' with a status of 'Succeeded'. A progress bar indicates 100% completion. Below the progress bar, there are tabs for Overview, Subagents (23), Findings (41), and Activity. The 'Working now' section shows 'No active subagent work right now.' On the right, the 'Artifacts' section lists two files: 'report.pdf' (489.6 KB) and 'localized_report report.ru.pdf' (500.8 KB), both with download buttons. The top navigation bar includes a 'New project' button, language settings (RU, EN), and user information (Triadix Admin).

Control view

- Stage progress
- Master status
- Report artifacts
- Run parameters

Subagents

Parallel work is visible as named subagents with progress and completion state.

The screenshot displays the TRIADIX PLATFORM interface for a customer workspace named 'AI pentest'. The left sidebar contains navigation links: Dashboard, AI pentest (selected), Projects, Calendar, Attack Surface, Diagnostics, and Organization. The main content area shows the project details for 'tensor.ctfonly.fans', including a progress bar at 100%, 1h 13m elapsed time, and 41 findings. A progress timeline at the top indicates stages: Preparation, Recon, Scope review, Checks, and Report. Below this, the 'Subagents' section is active, showing a list of subagents with their completion status and progress bars. The 'Artifacts' section on the right lists files ready for download, including 'report.pdf' (489.6 KB) and 'localized_report' (500.8 KB).

Subagent Name	Status	Progress
pentest-aigoat-admin-escalation	completed	100%
pentest-aigoat-prompt-injection	completed	100%
pentest-aigoat-sourcemap	completed	100%
pentest-aigoat-tokens	complete	100%
pentest-gitlab-register		

Accountability

- Named workers
- Progress per branch
- Completed work separated
- Artifact panel nearby

Task timeline

Expanded subagent cards show checkpoints, summaries, and findings for a focused task.

The screenshot displays the TRIADIX Platform interface. On the left is a sidebar with 'ACCOUNT' and links for 'Messages', 'Notifications', and 'Profile'. The main area shows a task timeline for 'pentest-aigoat-tokens' (status: complete). Above this, three other tasks are listed: 'pentest-aigoat-admin-escalation' (completed), 'pentest-aigoat-prompt-injection' (completed), and 'pentest-aigoat-sourcemap' (completed), each with a 100% progress bar. The timeline for 'pentest-aigoat-tokens' includes a date 'Jul 03, 2026, 11:34 PM' and a list of events: 11:28 PM (Worker prompt is waiting for human approval), 11:31 PM (Worker approved and waiting for supervisor start), 11:32 PM (Starting token-based API enumeration on aigoat.ctfonly.fans), 11:32 PM (Bearer token auth confirmed working. Testing all 4 tokens against multiple endpoints.), and 11:34 PM (Completed token-based API enumeration: 4/4 tokens work, 7 findings documented (2 HIGH, 3 MEDIUM, 2 LOW)). Below the timeline is a 'Findings' section with 7 confirmed entries. The first finding is 'Any Authenticated User Can Change Global AI Defense Level' (High severity), with details on TARGET (POST /api/chat/defense-level), IMPACT (Weakens AI security for ALL users, enabling prompt injection attacks.), and REMEDIATION (Require admin privileges or higher authorization to change defense level.). The second finding is 'Shared Knowledge Base Writable by Any Authenticated User (RAG Poisoning)' (High severity).

ACCOUNT

- Messages
- Notifications
- Profile

pentest-aigoat-admin-escalation
completed 100% ▾

pentest-aigoat-prompt-injection
completed 100% ▾

pentest-aigoat-sourcemap
completed 100% ▾

pentest-aigoat-tokens
complete 100% ▴

Jul 03, 2026, 11:34 PM

- 11:28 PM • Worker prompt is waiting for human approval
- 11:31 PM • Worker approved and waiting for supervisor start
- 11:32 PM • Starting token-based API enumeration on aigoat.ctfonly.fans
- 11:32 PM • Bearer token auth confirmed working. Testing all 4 tokens against multiple endpoints.
- 11:34 PM • Completed token-based API enumeration: 4/4 tokens work, 7 findings documented (2 HIGH, 3 MEDIUM, 2 LOW)

Findings
7 confirmed entries - Jul 03, 2026, 11:34 PM

Any Authenticated User Can Change Global AI Defense Level High

TARGET
POST /api/chat/defense-level

IMPACT
Weakens AI security for ALL users, enabling prompt injection attacks.

REMEDIATION
Require admin privileges or higher authorization to change defense level.

pentest-aigoat-tokens

Shared Knowledge Base Writable by Any Authenticated User (RAG Poisoning) High

localized_report
report.ru.pdf
500.8 KB
Download

Inspectable work

- Approval checkpoint
- Runtime milestones
- Evidence summary
- Finding linkage

Findings

Issues are presented with severity, affected target, evidence, impact, and remediation.

The screenshot displays the 'Findings' tab in the TRIADIX Platform. The left sidebar contains 'Organization' and 'ACCOUNT' (Messages, Notifications, Profile). The top navigation bar includes 'Overview', 'Subagents', 'Findings', and 'Activity'. The main content area, titled 'Findings in this run', lists three findings, each with a 'High' severity label.

Findings in this run
Confirmed issues and new signals that need triage.

Admin Credentials Exposed in Chatbot Context High

TARGET
Chatbot (LLM02 / Sensitive Info Disclosure)

EVIDENCE
AttacksPage.jsx LLM02 lab confirms backend chatbot context contains admin/admin123 credentials at defense level L0

IMPACT
Prompt injection attack can extract admin credentials from chatbot, leading to full admin panel access

REMEDIATION
Never include real credentials in LLM context; use L2 defense level to block extraction; remove sensitive data from RAG context
pentest-aigoat-sourcemap

Admin Credentials Exposed in Chatbot Context (Source Code Confirmation) High

TARGET
AI Chatbot System Prompt / Context

EVIDENCE
Source map analysis of AttacksPage.jsx (LLM02 lab) confirms the chatbot's context contains admin/admin123 credentials, database path (/app/aigoat.db), and config.yml

IMPACT
Prompt injection can extract admin credentials from chatbot context, granting full admin panel access

REMEDIATION
Never embed real credentials in LLM context; use L2 defense level to block extraction; remove sensitive data from RAG context
pentest-aigoat-prompt-injection

Any Authenticated User Can Change Global AI Defense Level High

TARGET
POST /api/chat/defense-level

Artifacts
Files ready for download.

- report report.pdf 489.6 KB Download
- localized_report report.ru.pdf 500.8 KB Download

Triage output

- Severity labels
- Affected target
- Evidence summary
- Remediation guidance

Activity and audit

Control-plane events, approvals, artifacts, and agent state are traceable.

The screenshot displays the Triadix Platform interface for a customer workspace named 'AI pentest'. The left sidebar contains navigation links: TP Triadix Production, Dashboard, AI pentest (selected), Projects, Calendar, Attack Surface, Diagnostics, and Organization. The main content area shows the project details for 'tensor.ctfonly.fans', including a progress bar at 100%, 1h 13m elapsed time, and 41 findings. A workflow diagram shows the stages: Preparation, Recon, Scope review, Checks, and Report. The 'Activity' tab is selected, showing an 'Execution log' with control-plane sync events. The log table has columns for TIME, EVENT, AGENT, and STATE. The 'Artifacts' section on the right lists two files: 'report.pdf' (489.6 KB) and 'localized_report.pdf' (500.8 KB), both with download buttons.

Execution log
Control-plane sync events for this run.

Activity

TIME	EVENT	AGENT	STATE
Jul 04, 2026, 12:24 AM	Worker synced	recon-whois	Done
Jul 04, 2026, 12:24 AM	Worker synced	recon-vision	Done
Jul 04, 2026, 12:24 AM	Worker synced	recon-web-subdomains	Done
Jul 04, 2026, 12:24 AM	Worker synced	recon-tensor-app	Done
Jul 04, 2026, 12:24 AM	Worker synced	recon-subdomains	Done
Jul 04, 2026, 12:24 AM	Worker synced	recon-ports	Done
Jul 04, 2026, 12:24 AM	Worker synced	recon-static-enum	Done

Artifacts
Files ready for download.

- report.pdf (489.6 KB) [Download]
- localized_report.pdf (500.8 KB) [Download]

Traceability

- Run events
- Worker sync
- Approval records
- Artifacts and report events

From live evidence to a customer-ready report

The customer sees conclusions and remediation, not raw internal files.

Capture

Assessment evidence is collected as the run executes: observations, screenshots, tool output, and reviewer notes.

Triage

Signals are grouped, deduplicated, assigned confidence, and filtered before becoming findings.

Deliver

The final output is an executive summary, technical findings, remediation guidance, and downloadable artifacts.

Result: a clear customer handoff with evidence, impact, confidence, and remediation.

Current implementation status

The feature is implemented as a real Triadix domain and live-verified in production.

Jul 3

production acceptance

24h

extended runner timeout

23

subagents in sample run

EN/RU

interface and reports

UI

approval workflow

- Platform domain: GraphQL, gRPC, migrations, RBAC, frontend workspace, Playwright coverage
- Runner: Ralph engine, worker sync, report generation, archive upload, cancel/timeout handling
- Safety: worker approval, auto-approve toggle, resource limits, network egress guard
- Production checks: signed report downloads, worker isolation, public ingress restrictions

Adoption path

Start narrow, define policy, then scale runner capacity and coverage.

Pilot

Start with a narrow verified scope.

Policy

Define approval and risk rules.

Scale

Add runners and broaden coverage.

A practical rollout starts with one authorized scope, one approval policy, and measurable report quality.

From AI agents to accountable pentest delivery

Triadix AI Pentest packages scope control, multi-agent execution, evidence handling, and reporting into one managed workflow.



TRIADIX AI PENTEST